Hindawi Publishing Corporation
Education Research International
Volume 2013, Article ID 198920, 13 pages
http://dx.doi.org/10.1155/2013/198920

Research Article

The Assessment of Critical Thinking Critically Assessed in Higher Education: A Validation Study of the CCTT and the HCTA

An Verburgh,[1] Sigrid François,[1] Jan Elen,[1] and Rianne Janssen[2]

[1] *Centre for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2, P.O. Box 3773, 3000 Leuven, Belgium*
[2] *Centre for Educational Effectiveness and Evaluation, KU Leuven, Dekenstraat 2, P.O. Box 3773, 3000 Leuven, Belgium*

Correspondence should be addressed to An Verburgh; an.verburgh@ppw.kuleuven.be

Received 26 June 2013; Revised 11 September 2013; Accepted 16 September 2013

Academic Editor: Lieven Verschaffel

Although critical thinking (CT) is generally acknowledged as an important aim of higher education, no validated instrument to assess CT in Dutch is available. Moreover, most instruments are validated on a broad sample with people of diverse educational backgrounds. This possibly hampers the reliability of assessing effects of instructional interventions within educational programmes, where diversity is less. This study investigates the psychometric quality of a translation of the Cornell Critical Thinking Test (CCTT) and the Halpern Critical Thinking Assessment (HCTA) in a sample of Dutch speaking freshmen majoring in educational sciences. Results show a higher content validity and preference by students for the HCTA. The CCTT, however, takes less time to administer and score, which makes it easier to use the CCTT on a larger scale. Neither of the two tests shows a high overall reliability. The strength of the correlations between the constructed-response items and the forced-choice items of the HCTA with the CCTT calls for further research on the precise relation between CT skills and dispositions and the ability of the HCTA to assess both independently.

1. Introduction

The development of critical thinking (CT) is generally acknowledged as an important aim of higher education [1–4]. Higher education graduates should be able to make decisions based on a well-thought consideration of available information. Research show that students grow in their CT abilities during college [5–8], but growth is slow and limited [4, 9–11]. There is however a lack of validated tests for CT development in Dutch speaking university students. The goal of the present study is therefore twofold: to investigate the psychometric properties of two commonly used tests for CT in Flemish university students within one discipline and to assess their progress in CT using these two tests during one academic year. The results of the study are also valuable outside the Dutch language community because the study adds to the overall understanding of CT and its assessment difficulties. Moreover, the study is confined to students in one discipline in order to know the reliability of the instruments within more restricted populations. There is a demand of CT measures that are able to evaluate instructional interventions [12]. Such instructional interventions are mostly conducted within one discipline, and hence, instruments need to be reliable within a restricted population.

In the following, the concept of CT is described first; afterwards, current tests on CT are discussed. Finally, the purpose of the present study and its design are presented.

1.1. The Concept of CT. Despite the widespread agreement on the importance of the development of CT in students, agreement on its precise meaning is lacking [13]. The latter is exemplified by the variety of existing definitions on CT [14–17]. At least two different considerations of the conceptualisation of CT can be discerned: (1) considering CT as discipline-specific and/or discipline-general and (2) considering CT as a set of skills or as a combination of skills with a disposition to be a "critical thinker".

Concerning the first aspect, Moore [18] distinguishes between two opposed movements in CT: the generalist movement and the discipline-specific movement. For the generalist movement, with Ennis [19] as leading figure, CT is a set of cognitive abilities that can be taught independently of a specific content. The discipline-specific movement, with McPeck [20] as leading figure, considers CT to be dependent on the problem area or the discipline under consideration. He argues that what counts as an appropriate use of scepticism in one discipline or context might be inappropriate in another. However, during the last decade, the discussion between the two movements has become less prominent as most researchers agree that there are some general CT skills, which are applicable in various contexts, while familiarity with a discipline plays an important role too [21].

A second facet on which scholars differ in their conceptualisation of CT concerns the question whether CT is a set of skills or also a disposition [22]. CT skills refer to, among others, rules of formal logic, consideration of multiple perspectives, induction, and deduction [21, 22]. In a dispositional viewpoint, the motivation and the commitment of a person are included too, to see whether a person is able to recognize and willing to use the needed CT [23]. The dispositional viewpoint encompasses a more holistic view on CT, which stipulates that skills as well as other dispositional components together influence a person's CT performance [22].

1.2. Tests of CT. The diversity of conceptualisations of CT is mirrored in a diversity of available discipline-specific and discipline-general tests of CT [24]. Even within discipline-general instruments, test developers depart from different conceptualisations of CT or place a different emphasis on particular aspects of CT. To give a few examples, the Reasoning about Current Issues Test (RCI) [25] is based on the Reflective Judgement Model by King and Kitchener [26]. The Watson-Glaser Critical Thinking Appraisal [27] aims at measuring CT as it is defined by Glaser [28]:

> (1) an attitude of being disposed to consider in a thoughtful way the problems and subjects that come within the range of one's experience; (2) knowledge of the methods of logical inquiry and reasoning; and (3) some skill in applying those methods. Critical thinking calls for a persistent effort to examine any belief or supposed form of knowledge in the light of the evidence that supports it and the further conclusions to which it tends. (p. 5)

The Cornell Critical Thinking Test (CCTT) [29] is inspired by the Cornell/Illinois model of CT. A fourth example is the Halpern Critical Thinking Assessment (HCTA) [30, 31], which is based on Halpern's definition [23]. As a final example, the California Critical Thinking Disposition Inventory (CCTDI) [32] claims to measure the inclination or disposition towards CT, as defined by Facione [33].

In addition to the diversity in conceptualisations, a wide range of item formats is used in tests of CT. Commonly used instruments such as the CCTT [29] use a forced-choice question format. However, in recent literature [22, 24, 34], it is argued that a combination of forced-choice and constructed-response items is more suitable to measure CT because the constructed-response format allows better grasping of the dispositional aspect of CT. Unlike forced-choice questions, constructed-response questions enable us to infer the respondent's reasoning behind an answer. Furthermore, the use of forced-choice questions may only indicate whether a respondent can recognize a correct answer or not, but it does not contain information about spontaneous answers from that respondent. As Ku [22] argues, if a test is intended to measure dispositions as well as skills, the test ought to allow respondents to think spontaneously. The HCTA [30, 31] is a test that combines constructed-response and forced-choice items. Apart from the above mentioned item formats, still other formats have been used, such as interviews (e.g., the Reflective Judgement Interview [35]), essays (e.g., Ennis-Weir Critical Thinking Essay Test [36]), a combination of essays and multiple-choice questions (e.g., the Critical Thinking Assessment Battery [37]), and Likert-type statements (e.g., the Problem Solving Inventory [38]).

1.3. The Present Study. A validated instrument for assessing CT in Dutch language students in higher education is lacking. Merely translating would not be sufficient to guarantee a valid instrument as cross-cultural assessment of generic skills as CT appears to be difficult [39]. Moreover, most tests are validated on a broad population, while most CT interventions are focused on students of one programme, and instruments need to be valid for a population with less variability. Therefore, the present study investigates the psychometric qualities of two instruments for assessing CT in students in higher education in Flanders, which is the Dutch speaking part of Belgium. Two internationally used CT tests were selected and administered to a sample of freshmen (first-year students), majoring in educational sciences.

Three criteria were used for selecting the two instruments [16–40]. Firstly, the selected instrument had to measure CT ability irrespective of discipline-specific knowledge of students. Therefore, only discipline-general tests were considered.

Secondly, the underlying conception of CT of the selected instrument needed to fit with how CT is understood and taught in the field under consideration: higher education (HE) in Flanders. In accordance with Cook et al. [40], a definition was established in close cooperation with representatives of HE institutions in Flanders. This definition fits with how CT is understood and presumably taught in higher education in Flanders. The representatives agreed with the following shortened version of the definition of CT by Facione [33], which considers CT as a combination of skills, leading a judgement, and dispositions, described as characteristics of the critical thinker:

> We understand critical thinking to be purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based. . .. The ideal

> critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgements, willing to reconsider, clear about issues, orderly in complex matters, and diligent in seeking relevant information. (p. 2)

Finally, evidence of the psychometric quality of the instrument needed to be available.

According to these criteria, two Anglo-Saxon instruments were selected: the Cornell Critical Thinking Test-Level Z (CCTT) [36] and the Halpern Critical Thinking Assessment (HCTA) [30, 31]. Although comparable in format, the CCTT was preferred above the Watson-Glaser Critical Thinking Appraisal [27] because of a better match with the definition. The Reasoning about Current Issues Test [25], although interesting in format, was not selected for this study because the interpretation of the scores was unclear.

The present study investigated the psychometric quality of both tests on four important aspects [16, 40]: (1) reliability, (2) validity, (3) feasibility of the administration and scoring of the tests, and (4) attractiveness of the test for the envisaged respondents.

Estimating the reliability of a test aims at estimating how much of variability in the scores of the test can be attributed to errors in the measurement and how much to true scores [41]. There are different types of reliability. Here, the interrater reliability and the internal consistency are investigated [16]. When a test is scored manually—as is the case for the HCTA—it is important that the scores are independent of the rater. Interrater reliability refers to this criterion. Reliability also refers to the internal consistency of the test [41], which gives an indication whether the items of a test that measure the same general construct produce similar scores.

The validity of a test refers to the extent to which the test accurately measures what it is supposed to measure [41–43]. There are different types of validity of which we will assess three: content validity, construct validity, and criterion validity. Content validity concerns the degree to which the items in the test cover the domain of the construct being measured. Construct validity indicates the extent to which the variables of a test accurately measure the construct under consideration. It can be assessed at the level of the test by investigating the relation between the test scores and other tests, which is also called congruent validity. Construct validity can also be assessed at the level of the items of a test, by using factor analysis. Criterion validity can be defined as the ability of a test to make correct predictions. Therefore, it is also often referred to as predictive validity. In our study, both tests were used to assess the progress in CT for freshmen between the beginning and the end of the academic year. In fact, these data can be considered as part of the investigation of the criterion validity of both tests.

Finally, the feasibility of the test concerns the ease of the test to administer and analyse. The attractiveness of a test relates to the extent to which respondents like the test [39]. It is assumed that the more attractive respondents find the test, the more they will be willing to commit to taking the test.

2. Method

2.1. Instruments

2.1.1. Assessment of CT

HCTA. The HCTA [30, 31] is a recently developed discipline-general test which consists of 25 descriptions of daily-life situations. Each situation is offered twice to the respondents: a first time followed by an open-ended question, where students have to construct their own answer (constructed-response item) and a second time followed by a forced-choice question (forced-choice item). The forced-choice items have different formats: multiple-choice questions with one or with more than one correct answer; rating questions with a Likert-type or with a yes/no scale; and matching questions. The maximum score per item ranges from one to ten. The test aims at measuring five categories of CT. Each category is measured in five situations. When calculating a total test score, the contribution of each category differs: (1) hypothesis testing (24%), (2) verbal reasoning (12%), (3) analysis of arguments (21%), (4) use of likelihood and uncertainty (12%), and (5) decision making and problem solving (31%).

The HCTA results in thirteen different scores. Apart from the total score, there is a total score for the constructed-response items (constructed-response part) and the forced-choice items (forced-choice part). In addition, there are five subscores in each category, both for the constructed-response items and the forced-choice items.

For the present study, the five categories of the HCTA show a good correspondence with the first part of the developed definition of CT. By using the constructed-response questions, the HCTA claims to be able to measure CT dispositions [44]. Because the HCTA is a very recent instrument, data about the psychometric features are mainly limited to the test manual and to research in close cooperation with the author. According to the test manual, the HCTA has a high internal consistency (Cronbach's $\alpha = 0.88$ for the total score) [31]. The constructed-response part and the forced-choice part separately show high internal consistencies as well (resp., $\alpha = 0.84$ and $\alpha = 0.79$). The respondents in the sample reported in the manual have diverse educational backgrounds, and their age ranges from 18 to 72. Reliability analyses of translations into Chinese and Spanish found Cronbach's α ranging from 0.69 to 0.77 for the overall test and low reliabilities for the subscales ($\alpha = 0.34$ to 0.64) [44–46]. In these studies, the sample consisted of students of different years and of different disciplines.

Correlations between the constructed-response part and the forced-choice part indicate that both parts measure related but distinct constructs [31]. Factor analyses point in the direction of a ten-factor structure (the five categories with a distinction between constructed-response and forced-choice items) [31]. Evidence for criterion validity has been established by Butler et al. [47], who found that the HCTA predicts real-world outcomes of CT. In addition there are significant correlations between epistemological beliefs and the HCTA [45].

CCTT. The CCTT [29] is a discipline-general test, intended for strong students in upper secondary education, students in higher education, and adults. Developed in 1985, it is a widely used instrument for assessing CT. It aims at measuring five aspects of CT: deduction; semantics; observation and credibility of sources; induction; definition and assumption identification. Each aspect is measured in a separate section in the test, but induction is split into two sections, namely, on the use of induction in hypothesis testing and in planning experiments. The test contains 52 items, all of which are in a forced-choice format. Similar to the HCTA, the CCTT has a relatively good correspondence with the first part of the definition. On the other hand, the second part of the definition, which captures the concept of CT as a disposition, is lacking, because the instrument—as most of the other instruments—only uses multiple-choice questions.

Regarding its reliability, the CCTT's manual reports split-half reliabilities between $r = 0.49$ and 0.80 and Kuder-Richardson reliabilities between KR = 0.50 and 0.76 [29]. The respondents in the studies were mostly undergraduate students or graduate students, mostly within one discipline. Taking all studies together, the CCTT was evaluated in a broad range of different institutions of higher education. Erwin [16] reports internal consistency values of $\alpha = 0.58$ in a sample of freshmen and of $\alpha = 0.72$ in a sample of sophomores.

The content validity of the CCTT was assessed by the members of the Illinois Critical Thinking Project, who agreed that the items of the CCTT measure CT as defined by the authors [29]. In addition, there are positive indications for criterion validity. For example, the correlation with the reflective Judgement Interview of King and Kitchener is 0.46 [48].

Translation Procedures. Both tests [29–31] were translated into Dutch, following the International Test Commission guidelines for translating and adapting tests [49]. This translation process consisted of several steps, following Wang et al. [50]. After a first translation was made, a pilot study ($N = 5$) was conducted, in which respondents filled out a translated test and were asked about their comments on the items of the tests during cognitive interviews. Adaptations to the translations were made. Next, a validation of the translation was done using the translation-back technique [51]. In this procedure, the Dutch versions were back translated in English by a third person and the translated version was compared with the original English version. The differences in both versions were discussed, and adaptations were made to the Dutch translations. Finally, both Dutch versions of the tests were administered to two different try-out groups of students in order to fine-tune the translation and cultural adaptation ($N = 66$ for the CCTT; $N = 40$ for the HCTA).

In order to establish cultural appropriateness for our population [50], several items of the HCTA were slightly changed during the translation process compared to the original English version. For example, one situation concerns a presidential candidate. This was changed into "politician", because the investigated population does not have a president. In another situation, respondents have to estimate the chances of a young woman to become a famous actress in Hollywood. The Dutch translation specifies that the young woman is American, because the tryout showed that some students' answers were influenced by the assumption that the woman came from a small, non-English speaking country, and that this lowered her chances. Although this was a correct inference, this answer was not intended by the original test.

For the CCTT, one item (item 21) was removed from the translated version, because this item is mainly based on the double meaning of the word "drugs". In Dutch there is no equivalent with the same double meaning.

2.1.2. Assessment of Attractiveness. The attractiveness of the tests was measured with a self-developed questionnaire consisting of three parts. In the first part of the questionnaire, students were asked to evaluate each test separately on a seven-point response scale (ranging from totally disagree to totally agree) concerning its difficulty, attractiveness, time, and amount of reading/writing necessary to fill in the test. For the CCTT, the use of the forced-choice questions only was also evaluated. Next, students had to make a forced choice between the two instruments, regarding features as being interesting, difficult, and of a good quality to show thinking ability. Finally, students could freely comment on the tests and explain their preference.

2.2. Procedure. Both tests were administered twice as a compulsory part of a first-year module. The tests were first administered at the beginning of the first semester (November) and again at the end of the second semester (May). The CCTT was administered on paper; the HCTA was online. For each test there were different collective sessions, from which students could choose their most convenient moment. In November, the collective sessions for the HCTA and the CCTT were mixed. In May, all the CCTT sessions were planned before the HCTA sessions. After finishing the HCTA in May, students were asked to fill in the attractiveness questionnaire. Each session started collectively. When students had finished, they could leave the room.

The answers to the CCTT and to the forced-choice questions of the HCTA were scored with the key of the manual. The answers on the constructed-response questions of the HCTA were scored according to the Vienna Test System [31], accompanied with the manual with examples. The Vienna Test System guides the rater through the respondent's answers on the constructed-response questions, with a series of prompts to be answered with "yes" or "no" or "yes", "no", or somewhat. This system is intended to increase the speed and the reliability of the scoring. After the establishment of interrater reliability (see Section 3.2.1 for details), the questions were scored by one rater.

In between the two administration moments, a workshop on CT with twenty representatives of different HE institutions was held. These representatives were partly different from the persons who developed the definition of CT. They were first asked to individually envisage a person in their own field who thinks critically and to write down what the person does. Then, they had to compare in small groups the activities

Table 1: Number of participants in each administration period, by gender.

Test	Gender	November	May
HCTA	Female	162	149
	Male	10	10
	Total	**172**	**159**
CCTT	Female	163	152
	Male	11	11
	Total	**174**	**163**

they wrote down and label the activities in abstract words. Afterwards, in a plenary session, they compared the activities with the used definition and with both instruments.

2.3. Participants. The participants were freshmen majoring in educational sciences at the KU Leuven (mean age = 18.2), Flanders, Belgium. In total 179 students filled out at least one test, of which 154 filled out both tests twice. The majority of the respondents were female, which is a normal situation for educational sciences in Flanders (see Table 1).

2.4. Analyses

2.4.1. Reliability. For the constructed-response questions of the HCTA, interrater reliability was investigated. Two raters individually scored the responses of 20 students. Afterwards, the differences were discussed in order to make sure that the prompts of the Vienna Testing System were understood in the same way. Next, the responses of 50 randomly selected students were scored by two raters. The results were compared, and differences were discussed. For each item, the proportion of equal scores and the weighted Cohen's κ were calculated. It was decided beforehand that if questions had a proportion of equal scores lower than 0.7, the responses of 50 additional students would be scored. During the establishment of interrater reliability, the Dutch version of the scoring guide was elaborated with more examples to ease the scoring and make it more transparent.

In order to make comparisons with the interrater reliabilities reported in the manual [31], the correlations between the subscale scores of the two raters were looked at. In addition, the effect of the rater on the means of the subscale scores was calculated by using a paired samples t-test.

The internal consistency was measured using the Cronbach's α. For both tests, it was calculated separately for the November and May administrations. For the HCTA, it was calculated for the overall result, for the constructed-response part, for the forced-choice part and for the five categories. Because in the HCTA the maximum score differs per item, two types of Cronbach's α were calculated: the normal Cronbach's α and the Cronbach's α of the standardized items. In addition to the analysis based on the scores of the individual items, the internal consistency was also calculated using the sum of the items of each subscale as a variable, because this approach was followed in the manual of the HCTA [31]. This additional calculation allows comparing our results with the results in the manual.

For the CCTT, Cronbach's α of sections IV and V and of sections VI and VII was calculated for both sections together, because these sections each measure the same or highly comparable aspects of the CCTT (resp., induction and assumption identification).

2.4.2. Validity. In order to assess the content validity we evaluated how every single aspect of the developed definition on CT was covered within both tests. These aspects of the developed definition are (1) purposeful, self-regulatory judgement, (2) interpretation, (3) analysis, (4) evaluation, (5) inference, (6) explanation of evidential and conceptual considerations, (7) explanation of methodological considerations, (8) explanation of criteriological considerations (9) explanation of contextual considerations, and (10) the "ideal critical thinker". Additional information on a close match between the conceptualization of CT in Flanders and the instruments was gained during the workshop on CT with twenty representatives of different HE institutions.

In the present study, the correlation between both tests in the same administration period can be used as an indication of construct validity. In addition to the observed correlation, also the correlation with correction for attenuation is used [52]. This correction allows correcting for a lack of perfect reliability, due to measurement errors which are inherent to empirical measures. Due to these measurement errors, the observed correlation is lower than the true correlation [53]. For the HCTA, the total score is taken into consideration, as well as the constructed-response part and the forced-choice part separately. It is expected that the CCTT will correlate higher with the total score and the forced-choice part of the HCTA than with the constructed-response part, because the constructed-response part is intended to measure also the dispositional aspect of CT, whereas the others are more restricted to CT skills.

In addition, the correlation between the constructed-response and the forced-choice part of the HCTA during the same administration period was looked at. Again, both the observed correlation and the correlation with correction for attenuation were considered. Because the two parts both measure aspects of CT, but with a different focus, a moderately strong correlation is expected.

In order to assess construct validity of the HCTA at item level, a principal component analysis (PCA) with an oblique rotation was planned on the 50 items for both November and May data. Based on the logic of the test, either five or two interdependent factors are expected (reflecting the skills and disposition measured in the respective forced-choices items and the constructed-response items). Before the analysis, the Kaiser-Meyer-Olkin (KMO) measure was used to verify the sample adequacy for a PCA. The latter is confirmed when KMO > 0.5 [54].

In order to assess the dimensionality of the CCTT, a Multidimensional Item Response Theory (MIRT) model was used [e.g., [55]]. A MIRT model is also called "item factor analysis". It is similar to a classical factor analysis in that it tries to assess the underlying dimensionality of a test. However, a MIRT model models the data set of the person by item responses directly, whereas a classical factor analysis models

Table 2: Descriptive statistics for the HCTA ($N = 155$).

Scale / Subscale	Max value	November				May				t	df	P
		M	SD	Min	Max	M	SD	Min	Max			
Total	195	116.08	10.11	84	140	120.32	10.88	92	147	5.144	154	.000
Constructed-response items (C)	95	49.99	6.68	31	66	52.81	7.40	37	68	4.885	154	.000
Forced-choice items (F)	99	66.09	5.33	52	77	67.52	5.73	53	83	3.058	154	.003
Hypothesis testing	45	26.35	4.14	16	38	27.43	3.83	17	36	2.990	154	.003
Hypothesis testing-C	18	9.49	2.42	3	16	10.54	2.44	3	16	4.607	154	.000
Hypothesis testing-F	27	16.86	2.75	9	23	16.90	2.66	11	23	0.130	154	.897
Verbal reasoning	22	11.26	2.49	5	18	11.52	2.53	4	18	1.177	154	.241
Verbal reasoning-C	15	7.06	2.32	2	14	7.20	2.25	2	13	0.668	154	.505
Verbal reasoning-F	7	4.20	0.86	2	6	4.32	0.94	1	7	1.364	154	.174
Argument analysis	42	25.10	4.45	14	35	26.88	4.30	13	36	4.360	154	.000
Argument analysis-C	23	12.10	2.94	5	19	12.81	3.17	4	20	−0.621	154	.535
Argument analysis-F	19	12.99	2.42	6	18	14.08	2.20	7	19	4.609	154	.000
Likelihood and uncertainty	24	13.96	2.99	4	21	14.75	3.10	6	21	3.115	154	.002
Likelihood and uncertainty-C	17	9.30	2.50	1	15	9.88	2.77	2	15	2.488	154	.014
Likelihood and uncertainty-F	7	4.66	1.08	2	7	4.87	0.90	2	7	2.237	154	.027
Decision making and problem solving skills	61	39.41	4.18	26	49	39.74	4.27	30	51	0.838	154	.403
Decision making and problem solving skills-C	22	12.04	2.81	6	18	12.39	2.70	6	20	1.295	154	.197
Decision making and problem solving skills-F	39	27.37	2.54	20	33	27.35	3.01	20	34	−0.072	154	.943

Table 3: Descriptive statistics for the CCTT ($N = 157$).

Scale	Max value	November				May				t	df	P
		M	SD	Min	Max	M	SD	Min	Max			
Total	51	27.13	4.24	15	37	27.53	4.39	16	39	1.049	156	.296
I Deduction	10	6.26	1.37	3	10	6.11	1.51	2	9	−0.947	156	.345
II Meaning & fallacies	10	3.45	1.49	0	8	3.95	1.49	1	8	3.419	156	.001
III Observation & credibility of sources	4	2.31	1.00	0	4	2.38	1.04	0	4	0.710	156	.479
IV/V Induction	17	9.50	1.91	5	14	9.32	1.83	4	14	−1.068	156	.287
VI/VII Definition & assumption identification	10	5.60	1.64	1	9	5.76	1.69	1	9	1.058	156	.292

the correlations over persons between the responses on the items of a test. Given this different approach, MIRT models are more apt to derive the dimensionality of a test with dichotomous items, because for such items, the correlation matrix is more difficult to assess [55]. The MIRT model was estimated using the R package mirt [56]. Models with one to five dimensions were compared using the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [57, 58]. The preference of one model above the other depends on the distance of the model with the data. The smaller the distance (the smaller the value of the criterion), the better the fit between the model and the data [59]. The best fit is expected for a model with five dimensions, given the five aspects of CT underlying the CCTT.

Criterion validity was assessed by looking at the correlation between the scores in November and May on the same test and by calculating the progress of individual students across both assessments.

2.4.3. Feasibility. The time to administer and to score the test was considered as criteria to assess the feasibility.

2.4.4. Attractiveness. With paired samples t-tests the appreciation of both tests was compared. In addition, the proportion of students preferring one test above the other was considered.

3. Results

3.1. Descriptive Statistics. Tables 2 and 3 describe the CT performance in November and May on the HCTA and the CCTT, respectively. On average, the total score was 116.08 in November and 120.32 in May on the HCTA. The difficulty level of the items varied. Some items were very easy (e.g., in May almost all students answered the forced-choice question of situation 9 correctly). Other items were difficult (e.g., in November almost 9 out of 10 students scored no points on the forced-choice question of situation 17). Similarly the subscales differed in difficulty: items of the argument analysis forced-choice subscale seemed easier for students than items of the hypothesis testing constructed-response subscale. The average difficulty varied between 0.15 and 0.99 proportions of correct answers.

Table 4: Interrater reliabilities for the constructed-response items of the HCTA ($N = 50$).

Situation	Proportion equal scores	Weighted Cohen's kappa
Situation 1	0.86	0.77
Situation 2	0.70	0.50
Situation 3	0.58 (0.72)*	(0.77)
Situation 4	0.88	0.85
Situation 5	0.72	0.62
Situation 6	0.56 (0.76)	(0.72)
Situation 7	0.84	0.78
Situation 8	0.86	0.77
Situation 9	0.86	0.76
Situation 10	0.94	0.83
Situation 11	0.90	0.89
Situation 12	0.86	0.90
Situation 13	0.70	0.58
Situation 14	0.92	0.90
Situation 15	0.66 (0.68)	(0.62)
Situation 16	0.78	0.72
Situation 17	0.88	/
Situation 18	0.74	0.74
Situation 19	0.72	0.75
Situation 20	0.88	0.75
Situation 21	0.74	0.73
Situation 22	0.80	0.71
Situation 23	0.70	0.68
Situation 24	0.98	0.94
Situation 25	0.66 (0.74)	(0.65)

Note: For question 17, it was impossible to calculate the weighted Cohen's kappa, because not all values were present. *The results between brackets are the results of a second set of 50 responses.

The average score on the CCTT was 27.13 in November and 27.53 in May. The difficulty levels of the items differed. Some items were hardly answered correctly (proportion of correct answers of 0.07) while others were almost always answered correctly (proportion of correct answers of 0.96). The difficulty of the subscales also differed: students answered more items correctly on the deduction scale than they did on the meaning and fallacies scale.

3.2. Reliability

3.2.1. Interrater Reliability. At item level, there was a high proportion of equal scores and of those satisfying weighted Cohen's κ, except for four situations (situations 3, 6, 15, and 25), as can be seen in Table 4. For these four items, the responses of 50 additional students were scored, and then, the results were satisfactory (indicated between brackets).

Table 5 shows the correlations between the scores of both raters on the five subscales. According to Cohen [60], these correlations were large. They were comparable or larger than those reported in the manual. However, the paired sample t-tests revealed—in contrast with the manual—a significant effect of rater for the constructed-response part, with a small effect size. For the subscales hypothesis testing and Likelihood and uncertainty, there was also a significant effect with a medium effect size. The scatter plots revealed that on these scales one rater systematically scored somewhat higher than the other.

3.2.2. Internal Consistency. Table 6 presents the internal consistency of the HCTA and the CCTT for the overall test and the different scales for the two test administrations. There was no much difference between both types of Cronbach's α. Therefore, the values of the standard Cronbach's α are reported below.

The November administration of the HCTA had an alpha of 0.53, which is a moderate internal consistency [60]. The internal consistencies of the constructed-response and forced-choice part separately were low (resp., $\alpha = 0.34$ and 0.35) as well as for the five categories of CT ($\alpha < 0.4$).

The internal consistency of the May administration of the HCTA was 0.64, which is acceptable [60]. In contrast to the November data, the constructed-response part and the forced-choice part had a moderate internal consistency (resp., $\alpha = 0.53$ and $\alpha = 0.49$). The internal consistencies of the separate CT categories of the HCTA were somewhat higher than in November, but still low to moderate ($\alpha < 0.43$).

The internal consistencies of the categories were also low in the two test administrations ($\alpha < 0.40$). Because of the low number of items in the subscales, the interitem correlations were considered, but these were also low (predominantly $r < 0.20$). For each scale, there was a limited number of items which correlated sufficiently with the total scale ($r > 0.25$) [61]. In addition, for each scale there were items that correlate negatively—but close to zero—with the scale total. The items with a sufficient correlation and with negative correlations differed between the two administration moments. Items which were very easy or difficult generally had a low correlation with the test total.

When calculating the internal consistency similarly as in the manual (taking the scores of the subscales as items), the α's were still low, with the exception of the overall score in May (Table 7).

For the CCTT, the internal consistency was moderate in November ($\alpha = 0.52$) (Table 6). Only two items correlated sufficiently with the total scale ($r > 0.25$), and one item correlated negatively with the total scale. The Cronbach's α's for the CCTT subscales were low ($\alpha < 0.30$). The internal consistency based on May data was the same as for November data ($\alpha = 0.52$), with five items sufficiently correlating with the total scale ($r > 0.25$) and seven items with a negative correlation with the total scale. Again, the α for the subscales was low ($\alpha < 0.30$), with the exception of assumption identification, which was somewhat higher ($\alpha = 0.42$).

3.3. Validity

3.3.1. Content Validity. Table 8 shows the results of the comparison between both tests and the developed definition. Both tests adequately mirrored the part of the definition on CT skills. The second part, with the dispositional aspects of CT, was measured by the constructed-response items of

Table 5: Interrater correlations and differences between the means between two raters on the constructed-response subscales of the HCTA.

Subscale—constructed-response items	Interrater correlations		Differences between the means of two raters			
	r in the sample	r in the manual	t	df	P	Cohen's d
Thinking as hypothesis testing	.85	.75	4.632	49	.000	0.37
Verbal reasoning	.84	.60	−0.414	49	.681	−0.03
Argument analysis	.88	.70	−1.531	49	.132	−0.10
Likelihood and uncertainty	.89	.82	3.357	49	.002	0.23
Decision making and problem solving	.84	.53	1.014	49	.315	0.08
Constructed-response part	.93	.83	3.063	49	.004	0.16

Table 6: Internal consistencies (Cronbach's α).

Scale	November		May	
	α	Standardized α	α	Standardized α
HCTA				
Total	0.53	0.55	0.64	0.64
Constructed-response part (C)	0.34	0.37	0.53	0.53
Forced-choice part (F)	0.35	0.37	0.48	0.44
Hypothesis testing	0.42	0.46	0.39	0.44
Verbal reasoning	0.21	0.17	0.28	0.28
Argument analysis	0.37	0.38	0.42	0.44
Likelihood and uncertainty	0.18	0.31	0.31	0.34
Decision making and problem solving	0.26	0.25	0.35	0.32
CCTT				
Total	0.52		0.52	
I Deduction	0.07		0.17	
II Meaning & fallacies	0.30		0.14	
III Observation & credibility of sources	0.17		0.25	
IV/V Induction	0.22		0.15	
VI/VII Definition & assumption identification	0.32		0.42	

Table 7: Internal consistency of the HCTA with scales as variable.

Scale	Sample		Manual
	November	May	
Total	0.50	0.58	0.88
Constructed-response part	0.28	0.46	0.84
Forced-choice part	0.32	0.40	0.79

the HCTA, where respondents had to formulate their own answers. This was not the case with the CCTT.

During the workshop with the representatives of different HE institutions, a close match between the perception of CT in Flanders' HE on the one hand and both tests on the other hand was found. All CT activities they identified were covered with the intended categories or sections of both tests.

3.3.2. Construct Validity. Table 9 shows the correlations between both tests in November in the upper left corner and in May in the lower right corner. All correlations were significantly different from zero, with a small strength [60]. The correlations corrected for attenuation indicated a relationship of medium strength between both tests.

In November as well as in May, the correlation between the constructed-response and the forced-choice parts of the HCTA was significantly different from zero, with a medium strength [60]. The correlation corrected for attenuation indicated a strong relation between both parts.

The Kaiser-Meyer-Olkin measure to verify the sample adequacy for a PCA indicated that November sample was inadequate, KMO = 0.45 [54]. The Kaiser-Meyer-Olkin measure for May sample was slightly better but still under the criterion of 0.5, KMO = 0.49. Therefore, it was decided to skip the analyses.

The fit statistics of the five exploratory factor solutions indicated that the unidimensional solution was the most parsimonious (Table 10). The AIC and BIC values were the lowest for the solution with one dimension, except for the AIC in November, where the two-dimension solution was slightly lower than the one-dimension solution.

3.3.3. Criterion Validity. The correlations between November and May results on the same (part of the) test were significantly different from zero (lower part of Table 9). When the correlations were corrected for attenuation, there was an indication of a strong relationship between the two test moments.

Table 8: Content validity: match between the used definition of CT and the elements measured in the two tests.

Aspects of CT	HCTA	CCTT
Skills		
Purposeful, self-regulatory judgment	All	All
Interpretation	Verbal reasoning	VI Identification of definitions and assumptions VII Identification of assumptions
Analysis	Argument analysis + situation 1 and 2	I Deduction
Evaluation	Verbal reasoning	II Meaning and fallacies
Inference	Argument analysis + situation 1 and 2	I Deduction
Explanation of evidential and conceptual considerations	Verbal reasoning	VI Identification of definitions and assumptions VII Identification of assumptions
Explanation of methodological considerations	Hypothesis testing Likelihood and uncertainty	III Observation and credibility of sources IV Induction (hypothesis testing) V Induction (planning experiments)
Explanation of criteriological considerations	Decision making and problem solving Likelihood and uncertainty	II Meaning and fallacies
Explanation of contextual considerations	Decision making and problem solving Likelihood and uncertainty	IV Induction (hypothesis testing) V Induction (planning experiments)
Disposition		
The ideal critical thinker	Constructed-response items	/

Table 9: Correlation between both tests in November and in May (with correction for attenuation).

	1	2	3	4	5	6	7	8
November								
(1) CCTT								
(2) HCTA total[1]	0.25** (0.35)							
(3) HCTA C	0.21** (0.36)							
(4) HCTA F	0.22** (0.37)		0.41** (0.70)					
May								
(5) CCTT	0.37** (0.51)	0.36** (0.49)	0.32** (0.44)	0.28** (0.39)				
(6) HCTA total	0.18* (0.22)	0.52** (0.66)	0.43** (0.54)	0.46** (0.57)	0.34** (0.42)			
(7) HCTA C	0.18* (0.25)	0.49** (0.67)	0.49** (0.67)	0.32 (0.44)	0.33** (0.45)			
(8) HCTA F	0.11 (0.15)	0.36** (0.52)	0.18* (0.27)	0.45** (0.65)	0.22** (0.32)		0.36** (0.52)	

*Significant at 0.05 level (2-tailed), **Significant at 0.01 level (2-tailed). [1]The correlation of the total of the HCTA with the constructed-response part and the forced-choice part is left out of the analysis, because the total is partly composed of each part.

Table 10: Fit statistics for MIRT models with one to five dimensions, for November and May administration of the HCTA.

Dimensions	AIC		BIC	
	November	May	November	May
1	9626	9113	9948	9429
2	9617	9113	10098	9583
3	9644	9139	10279	9761
4	9707	9175	10493	9945
5	9765	9249	10700	10164

Table 11: Students' opinions on the HCTA and the CCTT ($N = 132$).

Question	HCTA		CCTT		t	df	P
	M	SD	M	SD			
Test is difficult	4.84	1.13	5.52	.94	5,956	131	.000
Test is too difficult	3.20	1.39	4.11	1.56	5,472	130	.000
Test is fascinating	4.54	1.26	3.49	1.28	7,418	127	.000
Takes too long to fill in	5.05	1.50	4.41	1.55	3,832	131	.000
Too much reading to do my best	4.14	1.63	4.34	1.66	1,411	130	.161
Too much writing to do my best	3.12	1.55	/	/			
Would have done better in case of constructed-response opportunity	/	/	3.26	1.73			

For both tests, differences between both moments were investigated with paired sample t-tests. For the HCTA, there was a significant growth. This growth was confined to specific subcategories. Students did not advance in verbal reasoning and decision making and problem solving skills. They did advance neither on the hypothesis testing forced-choice subscale nor on the Argument analysis constructed-response subscale. For the CCTT, there was no growth.

3.4. Feasibility. The average testing time of the CCTT was 54 minutes. Students were rather neutral in their opinion when asked if the CCTT takes too long to complete (Table 11). About half of the students at least slightly agreed that the CCTT took too long. On average, the respondents needed about 80 minutes to complete the HCTA. When asked about their opinion on test administration time, on average students slightly agreed that it took too long to fill in the HCTA. The difference in opinion about test completion time between both instruments was significant ($P < 0.001$).

The scores of the CCTT and the forced-choice items of the HCTA could be calculated automatically because of the question format. The scoring of the constructed-response items of the HCTA could not be done automatically and was time consuming. It required some practice in order to score systematically. Estimated scoring time after a short training period was 15 min per test.

3.5. Attractiveness. On average, students slightly agreed that both instruments were difficult, but they found the CCTT significantly more difficult (Table 11). Students slightly disagreed that the HCTA was too difficult, while they agreed more that the CCTT was too difficult. This difference was significant ($P < 0.001$). On average students slightly disagreed that the CCTT was fascinating, while being neutral about the HCTA, which was again of a significant difference ($P < 0.001$). Finally, for both tests, students were neutral about whether they could have done better on the test if less reading would have been involved. There was no difference between both instruments. They slightly disagreed that they could have done better on the HCTA if they had to write less. They slightly disagreed that they could have done better on the CCTT if they would have had the opportunity to construct an answer themselves instead of being forced to choose an answer from a list.

Table 12: Students' preference of one of the instruments (%) ($N = 132$).

Choice	CCTT	HCTA
Most interesting	22.2	77.8
Most difficult	83.9	16.1
Most fascinating	23.5	76.5
Most challenging	37.0	63.0
Most motivating (first time)	27.8	72.2
Most motivating (second time)	27.2	72.8
Best showing my thinking ability	39.2	60.8
Preferred test	25.5	74.5

Students were also asked about their test preference (Table 12). About three-quarters of the respondents preferred the HCTA above the CCTT ($P < 0.001$). The HCTA was found to be the most interesting, fascinating, and motivating test ($P < 0.001$). To a lesser degree ($P < 0.01$), the HCTA was also deemed to be the most challenging test and the test with the greatest possibilities to demonstrate their thinking ability. Most students found the CCTT the most difficult test ($P < 0.001$). According to our respondents, the HCTA owed its appeal to its use of more familiar, recognizable everyday situations and constructed-response items. These items gave students the opportunity to express their own opinion, what they highly appreciated. Opposed to the HCTA, the CCTT's situations were not familiar, and its questions were regarded to be abstract. Students frequently mentioned the time needed to complete the HCTA and the fact that they had to write a lot themselves as negative points of this test. Hence, it comes as no surprise that the short nature of the multiple response format of the CCTT was cited most often as its major advance.

4. Discussion

This study compared two internationally widely used instruments to measure CT on their reliability, validity, feasibility, and attractiveness for students in higher education: the HCTA and the CCTT.

4.1. Reliability. The interrater reliability established in the study was satisfying, as measured in the high proportions of

equal scores and good to very good weighted Cohen's κ's, for almost all items [62]. In addition, the correlations between subscale totals of the two raters were strong. They were higher or comparable to those reported in the test manual [31]. However, in contrast to the manual, there was a significant effect of the rater on the scores. This might be caused by the fact that one rater scored systematically a little higher than the other. Despite this effect of the rater, the interrater reliability was satisfying. It indicates objectivity in the test scores. It seems that grading with the guiding prompts was a valuable way to reach a high interrater reliability.

The internal consistencies found in this study were low. This was not affected by the different weights of the items for calculating the total score, as indicated with the highly similar internal consistency on the raw or standardized item responses. Even when taking into account that the concept of CT is rather complex, the reported reliability measures are not sufficient. Only the reliability of the total score of the HCTA approaches acceptance. The internal consistencies on the tests found in the test manuals and in literature could not be replicated for the Dutch translations of the tests in the investigated research group. The difference between the Cronbach's α in this study and other studies might be due to the effect of the restriction of range or differences in the breadth of the population under investigation [53]. Reliability is dependent on the variance: the higher the variance, the higher the reliability. When a diverse population is considered, the variance is higher than that in a more restricted population, which has a positive effect on the reliability [53]. The reliability of the HCTA, reported in the manual, was assessed in a sample of respondents of different age groups and with diverse educational backgrounds [31], while the investigated population in this study is restricted to freshmen of one major at one university. A comparison between the descriptive statistics of the standardization sample in the manual and those of our sample indicated that although the means were comparable, the variances in the manual were indeed larger than in the present study, supporting the restriction of range explanation. Furthermore, most other studies describing reliabilities of the HCTA and the CCTT used broader samples than freshmen in one discipline.

4.2. Validity. The content validity of both tests is sufficient. Although neither of the two instruments captures the whole concept, they both match closely to the definition used by higher education staff members in Flanders. The HCTA has an extra strength by using a combination of constructed-response questions and the forced-choice questions, because the latter makes it possible to measure the dispositional aspects of CT.

Concerning the construct validity, the correlations between both tests during the same assessment period indicated a weak relationship. When the correlations are corrected for attenuation, the relationships indicated a medium to strong relationship. This suggests that the CCTT and the HCTA are—at least partially—measuring the same constructs and that the lack of correlation is partly due to the lack of reliability. Although the HCTA had a higher content validity because of its intention to measure the dispositional component of CT, this difference was not reflected in differences in strength of the correlations. In May, the relationship with the CCTT was even stronger for the constructed-response part than for the forced-choice part, while theoretically the opposite could be expected. The observed and corrected correlations between the constructed-response and forced-choice part of the HCTA are medium to strong. The strength of the latter correlations in combination with the strength of the correlations with the CCTT urges further research on the extent to which the constructed-response items are able to assess CT dispositions separately from the CT skills. Such a study could compare the strengths of the relations between each part of the HCTA and dispositions associated with "critical thinkers", such as tolerance for ambiguity, openness, and conscientiousness.

Construct validity on the level of the items of the test could not be assessed or confirmed. The results of the HCTA were not suitable for a PCA. The MIRT model analyses for the CCTT suggested that a unidimensional model fits better than a five-dimensional model. This finding indicates that the five separate scales correlate sufficiently high to form a single dimension. The sample size of the current study was not sufficient to estimate a confirmatory model with five dimensions with a correlated factor structure. Another explanation for the better fit of the one-dimensional model above the five-dimensional model are the low reliabilities of the subscales of the test. The latter may also be an explanation why the results of the HCTA were not suitable for PCA.

The medium to strong correlations between November and May scores on the same test are a positive indication of the criterion validity of both tests. The results of May administration can be predicted based on November scores.

With the HCTA, a growth in CT was assessed, while no difference was found between both administrations of the CCTT, although there was sufficient possibility for growth. This limited growth is in line with other research on development in CT (e.g., [9, 10]).

4.3. Feasibility and Attractiveness. Concerning feasibility and attractiveness, both instruments have their own strengths. The CCTT surpasses the HCTA with regard to feasibility. It took considerably less time to administer and to score. This difference was mainly due to the use of constructed-response questions in the HCTA. With regard to the test administration, there were more students who thought that the HCTA took too long to fill in than students who thought that the CCTT took too long. This was linked to the amount of writing that students had to do, as students indicated the quantity of writing as a disadvantage of the HCTA. With regard to the scoring, the Vienna Testing System was practical, but nevertheless scoring remained more time consuming than the scoring of the forced-choice questions. On the other hand, the use of constructed-response questions was one of the main strengths of the HCTA when it comes to content validity.

With regard to the attractiveness of both instruments, the students expressed that they preferred the HTCA above the CCTT, because the situations described in the test were more related to their daily-life experiences and because they could

express their own thinking more. Students considered both tests as difficult, but not too difficult. However, 40 percent claimed that the CCTT was at least slightly too difficult, compared to only one quarter for the HCTA. This finding may explain partly why still 50 percent of the students thought that the CCTT took too long, although it took considerably less time to fill in than the HCTA.

Summarizing, the present study showed that none of the two developed translations of the instruments is sufficient in reliability and validity for freshmen in educational sciences. The fact that only freshmen of one major were assessed is a plausible explanation for the different results in comparison to previous research on both tests. Although this is a limitation of the study, it could also point out that the tests might not be suited to study effects of CT interventions within one programme. The translated instruments hold some promising features, but adaptations in order to increase reliability and construct validity are required.

Moreover, the current study indicated that the constructed-response items of the HCTA are both the most appealing and content relevant characteristic as well as the major challenge of the test. The strengths of the correlations between the CCTT and the constructed-response items and the forced-choice items of the HCTA call for additional research on the question about the precise relation between CT dispositions and skills. Such research could in general focus on the overall question whether dispositions and skills are separately measurable and in particular whether the HCTA is capable of assessing both independently.

References

[1] D. Bok, *Our Underachieving Colleges*, Princeton University Press, Princeton, NJ, USA, 2006.

[2] T. Moore, "Critical thinking: seven definitions in search of a concept," *Studies in Higher Education*, vol. 38, no. 4, pp. 506–522, 2013.

[3] J. Vandermensbrugghe, "The unbearable vagueness of critical thinking in the context of the Anglo-Saxonisation of education," *International Education Journal*, vol. 5, no. 3, pp. 417–422, 2004.

[4] P. K. Wood and C. Kardash, "Critical elements in the design and analysis of studies of epistemology," in *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*, B. K. Hofer and P. R. Pintrich, Eds., pp. 231–260, Lawrence Erlbaum, Mahwah, NJ, USA, 2002.

[5] A. W. Astin, *What Matters in College? Four Critical Years Revisited*, Jossey Bass, San Fransico, Calif, USA, 1993.

[6] A. Gellin, "The effect of undergraduate student involvement on critical thinking: a meta-analysis of the literature 1991–2000," *Journal of College Student Development*, vol. 44, no. 6, pp. 746–762, 2003.

[7] C. A. Giancarlo and P. Facione, "A look across four years at the disposition towards critical thinking among undergraduate students," *The Journal of General Education*, vol. 50, pp. 29–55, 2001.

[8] M. A. Miller, "Outcomes evaluation: measuring critical thinking," *Journal of Advanced Nursing*, vol. 17, no. 12, pp. 1401–1407, 1992.

[9] R. Arum and J. Roksa, *Academically Adrift: Limited Learning on College Campuses*, The University of Chicago Press, Chicago, Ill, USA, 2011.

[10] E. T. Pascarella and P. T. Terenzini, "Cognitive skills and intellectual growth," in *How College Affects Students. Volume 2: A Third Decade of Research*, pp. 155–212, Jessey-Bass, San Francisco, Calif, USA, 2005.

[11] E. T. Pascarella, C. Blaich, G. L. Martin, and J. M. Hanson, "How robust are the findings of academically adrift?" *Change*, vol. 43, pp. 20–24, 2011.

[12] J. S. Clifford, M. M. Boufal, and J. E. Kurtz, "Personality traits and critical thinking skills in college students: empirical tests of a two-factor theory," *Assessment*, vol. 11, no. 2, pp. 169–176, 2004.

[13] P. C. Abrami, R. M. Bernard, E. Borokhovski et al., "Instructional interventions affecting critical thinking skills and dispositions: a stage 1 meta-analysis," *Review of Educational Research*, vol. 78, no. 4, pp. 1102–1134, 2008.

[14] S. Bailin, R. Case, J. R. Coombs, and L. B. Daniels, "Conceptualizing critical thinking," *Journal of Curriculum Studies*, vol. 31, no. 3, pp. 285–302, 1999.

[15] H. Butler, "Halpern critical thinking assessment predicts real-world outcomes of critical thinking," *Applied Cognitive Psychology*, vol. 26, pp. 721–729, 2012.

[16] T. D. Erwin, *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving and Writing*, Government Printing Office, Washington, DC, USA, 2000.

[17] P. Facione, "Critical thinking: what is it and why it counts," 2010, http://www.insightassessment.com/pfd_files/what&why-2010.pdf.

[18] T. Moore, "The critical thinking debate: how general are general thinking skills?" *Higher Education Research & Development*, vol. 23, no. 1, pp. 3–18, 2004.

[19] R. H. Ennis, "The degree to which critical thinking is subject specific: clarification and needed research," *Educational Researcher*, vol. 18, pp. 4–10, 1992.

[20] H. McPeck, "Critical thinking and subject specificity: a reply to Ennis," *Educational Researcher*, vol. 19, pp. 10–12, 1990.

[21] C. Angeli and N. Valanides, "Instructional effects on critical thinking: performance on ill-defined issues," *Learning and Instruction*, vol. 19, no. 4, pp. 322–334, 2009.

[22] K. Y. L. Ku, "Assessing students' critical thinking performance: urging for measurements using multi-response format," *Thinking Skills and Creativity*, vol. 4, no. 1, pp. 70–76, 2009.

[23] D. F. Halpern, *Thought and Knowledge: An Introduction to Critical Thinking*, Lawrence Earlbaum, New Jersey, NJ, USA, 2003.

[24] C. Saiz and S. Rivas, "Assessment in critical thinking: a proposal for differentiating ways of thinking," *Ergo Nueva Epoca*, vol. 22-23, pp. 25–66, 2008.

[25] P. M. King, K. S. Kitchener, and P. K. Wood, "The reasoning about current issues test," 2000, http://www.reflectivejudgment.org/.

[26] P. M. King and K. S. Kitchener, "The reflective judgement model: twenty years of research on epistemic cognition," in *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*, B. K. Hofer and P. R. Pintrich, Eds., pp. 37–61, Lawrence Erlbaum, Mahwah, NJ, USA, 2002.

[27] G. Watson and E. M. Glaser, *Watson-Glaser Critical Thinking Appraisal*, Psychological Corporation, Cleveland, Ohio, USA, 1980.

[28] E. Glaser, *An Experiment in the Development of Critical Thinking*, Bureau of Publications, Teacher College, Columbia University, New York, NY, USA, 1941.

[29] R. H. Ennis, J. Millman, and T. N. Tomko, *Cornell Critical Thinking Test*, Midwest, Pacific Grove, Calif, USA, 5th edition, 2005.

[30] D. F. Halpern, *Halpern Critical Thinking Assessment Using Everyday Situations: Background and Scoring Standards*, Claremont McKenna College, Claremont, Calif, USA, 2007.

[31] D. F. Halpern, *Halpern Critical Thinking Assessment: Manual, Version 22*, Schufried, Mödling, Austria, 2012.

[32] P. Facione and N. Facione, *The California Critical Thinking Disposition Inventory*, California Academic Press, Milbrae, Calif, USA, 1992.

[33] P. Facione, *Critical Thinking: A Statement of Expert Consensus of Purposes of Educational Assessment and Instruction*, California Academic Press, Millbrae, Calif, USA, 1990.

[34] S. Yeh, "Tests worth teaching to: constructing state-mandated tests that emphasize critical thinking," *Educational Researcher*, vol. 30, pp. 12–17, 2001.

[35] P. M. King and K. S. Kitchener, *Developing Reflective Judgement*, Jossey Bass, San Francisco, Calif, USA, 1994.

[36] R. H. Ennis and E. Weir, *The Ennis-Weir Critical Thinking Essay Test*, Midwest, Pacific Grove, Calif, USA, 1985.

[37] American College Testing Program, *CAAP Technical Handbook*, American College Testing Program, Iowa City, Iowa, USA, 1991.

[38] P. Heppner, *The Problem-Solving Inventory Manual*, Consulting Psychology Press, Palo Alto, Calif, USA, 1988.

[39] OECD, *Assessment of Higher Education Learning Outcomes AHELO: Feasibility Study Resport*, OECD, Paris, France, 2013.

[40] P. Cook, R. Johnson, P. Moore et al., "Critical thinking assessment: measuring a moving target," in *Report & Recommendations of the South Carolina Higher Education Assessment Network Critical Thinking Task Force*, The South Carolina Higher Education Assessment Network, Rock Hill, SC, USA, 1996.

[41] K. R. Murphy and C. O. Davidshofer, *Psychological Testing: Principles and Applications* , Prentice-Hall, Englewood Cliffs, NJ, USA, 2nd edition, 1991.

[42] P. Kline, *Handbook of Psychological Testing*, Routledge, London, UK, 1999.

[43] W. P. Vogt, *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*, Sage, Newbury Park, Calif, USA, 1999.

[44] K. Y. L. Ku and I. T. Ho, "Dispositional factors predicting Chinese students' critical thinking performance," *Personality and Individual Differences*, vol. 48, no. 1, pp. 54–58, 2010.

[45] N.-M. Chan, I. T. Ho, and K. Y. L. Ku, "Epistemic beliefs and critical thinking of Chinese students," *Learning and Individual Differences*, vol. 21, no. 1, pp. 67–77, 2011.

[46] A. M. Nieto, C. Saiz, and B. Orgaz, "Análisis de la propiedades psicométricas de la versión espanola del HCTAES-test de Halpern para la evaluaciòn del pensamiento crítico mediante situaciones cotidianas," *Revista Electrónica de Metodología Apliciada*, vol. 14, pp. 1–15, 2009.

[47] H. A. Butler, C. P. Dwyer, M. J. Hogan et al., "The Halpern critical thinking assessment and real-world outcomes: cross-national applications," *Thinking Skills and Creativity*, vol. 7, no. 2, pp. 112–121, 2012.

[48] P. M. King, P. K. Wood, and R. A. Mines, "Critical thinking among college and graduate students," *Review of Higher Education*, vol. 13, pp. 167–186, 1990.

[49] International Test Commission (2010), "*International Test Commission guidelines for translating and adapting tests*," http://www.intestcom.org.

[50] W.-L. Wang, H.-L. Lee, and S. J. Fetzer, "Challenges and strategies of instrument translation," *Western Journal of Nursing Research*, vol. 28, no. 3, pp. 310–321, 2006.

[51] W. Maneesriwongul and J. K. Dixon, "Instrument translation process: a methods review," *Journal of Advanced Nursing*, vol. 48, no. 2, pp. 175–186, 2004.

[52] P. M. Muchinsky, "The correction for attenuation," *Educational and Psychological Measurement*, vol. 56, no. 1, pp. 63–75, 1996.

[53] J. Scheerens, C. Glas, and S. M. Thomas, *Educational Evaluation, Assessment and Monitoring: A Systematic Approach*, Swets & Zeitlinger, Lisse, The Netherlands, 2003.

[54] A. Field, *Discovering Statistics Using SPSS (and Sex, Drugs and Rock "n" Roll)*, Sage, Newbury Park, Calif, USA, 2009.

[55] R. D. Bock, R. Gibbons, and E. Muraki, "Full-information item factor analysis," *Applied Psychological Measurement*, vol. 12, pp. 261–280, 1988.

[56] R. P. Chalmers, "mirt: a multidimensional item response theory package for the R environment," *Journal of Statistical Software*, vol. 48, pp. 1–29, 2012.

[57] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2002.

[58] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.

[59] H. Linhart and W. Zucchini, *Model Selection*, John Wiley & Sons, New York, NY, USA, 1986.

[60] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale, NJ, USA, 2nd edition, 1988.

[61] R. Ebel and D. Frisbie, *Essentials of Educational Measurement*, Prentice Hall, Englewood Cliffs, NJ, USA, 1991.

[62] U. Jakobsson and A. Westergren, "Statistical methods for assessing agreement for ordinal data," *Scandinavian Journal of Caring Sciences*, vol. 19, no. 4, pp. 427–431, 2005.